A Non-Asymptotic Analysis for Stein Variational Gradient Descent: Project report

Quentin Bourbon
MVA - Télécom Paris
quentin.bourbon@telecom-paris.fr

Matthieu Carreau
MVA - Télécom Paris
matthieu.carreau@telecom-paris.fr

Naël Farhan
MVA - ENSAE
nael.farhan@ensae.fr

Abstract

In this report, we present results established in [1] which lead to a study on the Stein Variational Gradient Descent (SVGD) first described in [3]. We also explore and illustrate properties shown in [1] with some basic experiments. We are also taking this opportunity to look at the behaviour of SVGD for more complex cases. The code of our implementation and experiments is available on this repository.

1 Introduction

Sampling points from a target distribution is a well-known problem in Bayesian inference. Some methods, such as Markov Chain Monte Carlo (MCMC), exist to solve this problem, but they are often slow or not scalable at all. In this context, Stein Variational Gradient Descent (SVGD) has been developed in [3] by considering the problem as an optimization problem in the Wasserstein space, aiming to minimize the KL divergence between a set of particles and the target distribution. The article under review [1] provides a non-asymptotic analysis for the SVGD algorithm when the distribution π admits a density proportional to $\exp(-V)$ where V is smooth enough.

After describing the work in [3] about SVGD in the first part, we will move on to [1] and its contribution to SVGD. In these two sections, we'll take the opportunity to illustrate using some basic experiments. After that, we will give some additional experiments which allow us to go further and see the limit of these articles [1] [3].

2 Stein Variational Gradient Descent

In this section, we provide the necessary background on SVGD to understand [1]; hence, we explain SVGD developed in [3]. Let's begin with some notation that we will use throughout the report. We will denote π as the target density and μ as the current density. Notations: Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel. We define the reproducing kernel Hilbert space $\mathcal{H}$ of k as the closure of Span $(\{\phi \mid \phi(x) = \sum_i a_i k(x, x_i), a_i \in \mathbb{R}, m \in \mathbb{N}, x_i \in \mathcal{X}\})$ equipped with the inner product $\langle \phi, \psi \rangle_{\mathcal{H}} = \sum_{ij} a_i b_j k(x_i, x_j)$ when $\phi(x) = \sum_i a_i k(x, x_i)$ and $\psi(x) = \sum_j b_j k(x, x_j)$.

Let $\mathcal{X} = \mathbb{R}^d$ be the space of distributions and $\mathcal{P}_2(\mathcal{X}) = \{\mu \in \mathcal{M}(\mathcal{X}) \mid \int ||x||^2 d\mu(x) < \infty\}$. Let $\mu, \pi \in \mathcal{P}_2(\mathcal{X})$. We define the *Kullback-Leibler divergence* between μ and π as:

$$\mathrm{KL}(\mu, \pi) = \mathbb{E}_\mu [\log \mu(x)] - \mathbb{E}_\mu [\log \pi(x)]$$

Samely, we define the *Stein operator* $\mathcal{A}_\pi$ of π as:

$$\forall \phi \in \mathcal{H}, \forall x \in \mathcal{X}, \mathcal{A}_\pi \phi(x) = \phi(x) \nabla_x \log \pi(x)^\top + \nabla_x \phi(x)$$

where $\mathcal{H}$ is a space of smooth functions.

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel. We define the reproducing kernel Hilbert space $\mathcal{H}$ of k as the closure of Span $(\{\phi \mid \phi(x) = \sum_i a_i k(x, x_i), a_i \in \mathbb{R}, m \in \mathbb{N}, x_i \in \mathcal{X}\})$ equipped with the inner product $\langle \phi, \psi \rangle_{\mathcal{H}} = \sum_{ij} a_i b_j k(x_i, x_j)$ when $\phi(x) = \sum_i a_i k(x, x_i)$ and $\psi(x) = \sum_j b_j k(x, x_j)$. Classically, we use the RBF kernel defined as $k(x, x') = \exp\left(-\frac{1}{h}||x - x'||_2^2\right)$.

For $T : \mathcal{X} \to \mathcal{Y}$, the push forward measure $\beta = T_{\#}\alpha \in \mathcal{M}(\mathcal{Y})$ of some $\alpha \in \mathcal{M}(\mathcal{X})$ satisfies

$$\forall h \in \mathcal{C}(\mathcal{Y}), \quad \int_{\mathcal{Y}} h(y)\, d\beta(y) = \int_{\mathcal{X}} h(T(x))\, d\alpha(x).$$

Equivalently, for any measurable set $B \subset \mathcal{Y}$, one has

$$\beta(B) = \alpha\left(\{x \in \mathcal{X} \mid T(x) \in B\}\right).$$

Problem Formulation The Stein's Identity states that for all $\phi \in \mathcal{H}$, we have

$$\mathbb{E}_{x\sim\pi}\left[\mathcal{A}_\pi \phi(x)\right] = 0$$

as soon as $\lim_{||x||\to\infty} \pi(x)\phi(x) = 0$.

Hence, the value of the magnitude of $\mathbb{E}_{x\sim\mu}\left[\mathcal{A}_\pi \phi(x)\right]$ relates to how different μ and π are. Therefore, one can define the *Stein discrepancy* as;

$$\mathbb{S}(\mu, \pi) = \max_{\phi\in\mathcal{H}} \mathbb{E}_{x\sim\mu}\left[\text{Tr}(\mathcal{A}_\pi \phi(x))\right]^2$$

Therefore, one wants to solve the following optimization problem

$$\min_{\mu} \mathbb{S}(\mu, \pi) = \min_{\mu} \max_{\phi\in\mathcal{H}} \mathbb{E}_{x\sim\mu}\left[\text{Tr}(\mathcal{A}_\pi \phi(x))\right]^2 \tag{1}$$

So far we have left aside the choice of $\mathcal{H}$. The idea developed in [3] is to emancipate ourselves from the challenging problem of maximising (1) over a too big space, by constraining ϕ to be in the unit ball. One can show that a solution to (1) over the unit ball is

$$\phi(x) = \frac{\phi^*_{\mu,\pi}(x)}{||\phi^*_{\mu,\pi}||^2_{\mathcal{H}^d}} \tag{2}$$

with $\phi^*_{\mu,\pi}(.) = \mathbb{E}_{x\sim\mu}\left[\mathcal{A}_\pi k(x, .)\right]$ for which we have $\mathbb{S}(\mu, \pi) = ||\phi^*_{\mu,\pi}(x)||^2_{\mathcal{H}^d}$.

SVGD The idea of [3] to minimize $\text{KL}(., \pi)$ is to consider it as a functional and then apply a gradient descent-like algorithm to it using the gradient of $\text{KL}(., \pi)$ under the inner product of $\mathcal{H}$

$$\mu_{t+1} = \left(I + \gamma_t \phi^*_{\mu_t,\pi}\right)_{\#} \mu_t \tag{3}$$

where γ_t is the step size at the t-th iteration and $\phi^*_{\mu_t,\pi}$ as defined in (2)

An important remark is that the gradient of $\text{KL}(., \pi)$ is composed of two terms. A term that drives the particles towards the high probability areas of $\pi(x)$ and a term that prevents all the points from collapsing together into local modes of $\pi(x)$.

Basic Experiments To highlight the main results of SVGD, we ran a few experiments on basic examples involving Gaussian and mixture of Gaussians in dimension 1. The setting was as follows: taking $n = 200$ points and running 1000 iterations. In figure 1, we ran SVGD on the following target distributions: one simple Gaussian with mean 2.5 and variance 0.1, and a mixture of three Gaussians. For a set number of points, SVGD approaches a simple Gaussian well but struggles with complexity, notably with the lower weighted part of the mixture as it seems to take a lot more iterations to correctly model the target.

3 Paper Summary

In this section, we present the contribution of [1] regarding SVGD. In their article, they assume that π has a density proportional to $\exp(-V)$, where V is smooth enough.

Figure 1: Basic Experiments in 1D.

Stein Fisher information For $\mu \in \mathcal{P}_2(\mathcal{X})$, the Stein Fisher Information of μ relative to π is defined by :

$$I_{Stein}(\mu|\pi) = \left\| S_\mu \nabla \log(\frac{\mu}{\pi}) \right\|^2_{\mathcal{H}}, \text{with } S_\mu \text{the Kernel integral operator: } S_\mu f = \int k(x,.)f(x)d\mu(x) \tag{4}$$

This is an adaptation of the Fisher information we know of in a parametric setting. Here, it's known as the squared Kernel Stein Discrepancy (KSD) providing discrepancy between distributions; note that we only need to know π up to a constant. When this converges to 0 i.e. $I_{Stein}(\mu_n|\pi) \to 0$, with some conditions on π (works for Gaussian mixtures), we have the weak convergence $\mu_n \to \pi$.

The article also introduces an important condition on π, called the *Stein log-Sobolev* inequality : π satisfies it with constant $\lambda > 0$ when $\text{KL}(\mu|\pi) < \frac{1}{2\lambda} I_{Stein}(\mu|\pi)$.

Inequalities They define the SVGD gradient flow as the solution of the following differential equation on μ_t:

$$\frac{\partial \mu^t}{\partial t} + \text{div}(u_t v_t) = 0; v_t = P_{\mu_t} \nabla \log(\frac{\mu_t}{\pi}) \tag{5}$$

Then, it proves an important result on the dissipation of the KL during SVGD in continuous time:

$$\frac{d\text{KL}(\mu_t|\pi)}{dt} = -I_{Stein}(\mu_t|\pi) \tag{6}$$

Under some regularity assumptions on the kernel and the target density, as long as $\mathbb{E}_{x\tilde{\mu}_t}(||x||)$ is bounded as a function of t, $\lim_{t\to\infty} I_{Stein}(\mu_t|\pi) = 0$.

Moreover, a convergence rate of $I_{Stein}(\mu_t|\pi)$ is obtained by integrating equation 6:

$$\min_{0\le s\le t} I_{Stein}(\mu_t|\pi) \le \frac{1}{t}\int_0^t I_{Stein}(\mu_s|\pi)ds \le \frac{\text{KL}(\mu_0|\pi)}{t} \tag{7}$$

When π satisfies the *Stein log-Sobolev* inequality with $\lambda > 0$, then it is shown that the objective function decreases exponentially:

$$\text{KL}(\mu_t||\pi) \le e^{-2t\lambda}\text{KL}(\mu_0||\pi) \tag{8}$$

SVGD in Discrete Time Their main contribution is to show similar descent lemmas in the discrete-time version of the SVGD.

In particular, they show that under some regularity assumptions, for a small enough step size γ and as long as the sequence of Stein-Fischer information is bounded, the KL divergence decreases at each iteration:

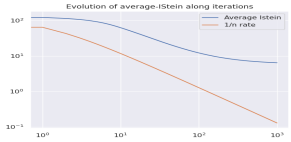
$$\text{KL}(\mu_{n+1}|\pi) - \text{KL}(\mu_n|\pi) \le -\gamma\left(1 - \gamma\frac{(\alpha^2 + M)B^2}{2}\right) I_{Stein}(\mu_n|\pi) \tag{9}$$

A very similar inequality was derived in [2], assuming the boundedness of the Kernel Stein Discrepancy. However, this article brings a new point of view by proving this proposition using differential calculus in the Wasserstein space, which allows us to replace this hypothesis with a weaker one, regarding the boundedness of the sequence $(I_{Stein}(\mu_n|\pi))_n$.

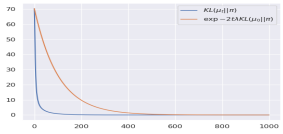
A consequence of this proposition is to obtain an equivalent of the inequality 7 for the discrete time algorithm:

$$\min_{k=1,\dots n} I_{Stein}(\mu_k|\pi) \leq \frac{1}{n}\sum_{k=1}^{n} I_{Stein}(\mu_k|\pi) \leq \frac{\mathrm{KL}(\mu_0|\pi)}{c_\gamma n} \tag{10}$$

In our basic experiments, we highlighted bounds on both the KL 8 and the Stein-Fisher information 9, as we can see from figure 2 for the 1D case with a simple Gaussian. Note that for Fisher, we didn't compute the exact formula for the constant c_γ, but we can still observe some kind of exponential decrease (linear in log-log scale).



(a) Average I_{Stein} in log-log scale



(b) Evolution of KL

Figure 2: Comparison of bounds in 1D.

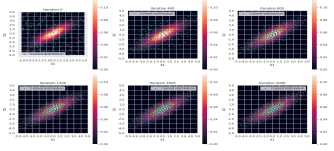
Finite Particle An important thing to note is that all of the bounds derived previously concern the sequence of continuous distributions $(\mu_n)_n$ for which we cannot compute KL divergences to π in practice. The point clouds $(\hat{\mu}_n)_n$ updated in the algorithm are only a set of samples which approximate these distributions. The final contribution of this paper is to quantify how well these point clouds represent these distributions in terms of Wasserstein distances, for any $0 \leq n \leq \frac{T}{\gamma}$,

$$\mathbb{E}[W_2^2(\mu_n, \hat{\mu}_n)] \leq \frac{1}{2}\left(\frac{1}{\sqrt{N}}\sqrt{\mathbb{V}\mathrm{ar}(\mu_0)}e^{2LT}\right)(e^{2LT}-1) \tag{11}$$

Where L is a constant depending on k and π.

4 Addtional Experiments

In this section, we will add some more complex experiments to observe the behavior of SVGD in different scenarios. To do so, we will compare the outcomes with those obtained with basic parameters. The target distribution is a 2D Gaussian distribution with $n = 200$ points and $h = 0.2$. We present the results in 3. We demonstrate that it converges very well in 2000 iterations.



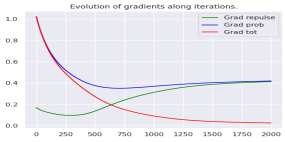
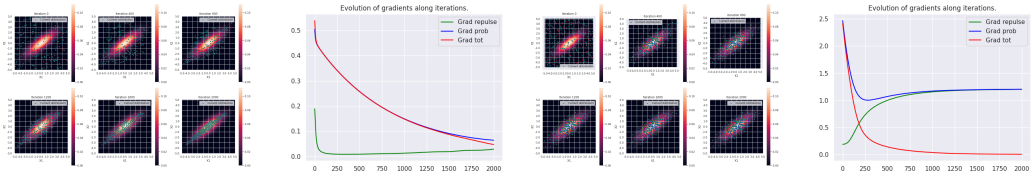


Figure 3: Result of the reference experiment.

Change of n Increasing the number of points takes a little longer to calculate (particularly for the kernel) but is still perfectly acceptable. On the other hand, convergence requires more operations, which also contributes to a longer calculation time. Indeed, we have tested this on a 2D Gaussian, going from $n = 200$ to $n = 600$, present in 4.

Change of Kernel We try to change the parameter h of the Gaussian kernel (which represents the variance to within one multiplicative factor). We try to reduce and then increase it. We show the result in 5. SVGD converges but differently. h seems to control how dense the points are at the π maxima. It's visible in the evolution of the gradients, we see that norm of the gradients (and especially the repulsive one) seems to increase with h as we can see in 5.

Figure 4: Comparison of experiments with high number of points.



(a) Experiment with low h.

(b) Experiment with high h.

Figure 5: Comparison of experiments with different h values.

Change of π We also try to sample from more complex distributions. Therefore, we try with a mixture of Gaussians in 2D and with a circular shape. We can see in 6 that SVGD works well but requires more iterations.

Figure 6: Comparison of complex target distributions.

5 Conclusion / Discussion

Finally, the article [1] under review provides strong theoretical results on SVGD that can be useful in practice (thanks to 11 that justifies using a finite number of particles) because it treats a wide range of target densities with weak assumptions.

However, all the results are shown with a very large bound that decreases slowly.

As a minor comment, we have also noticed a typographical error in assumption B1 of the published article, where V should be replaced by ∇V:

$$\exists C_V > 0, \forall x \in \mathcal{X}, ||\nabla V(x)|| < C_V$$

This hypothesis is actually quite strong as it is written and forbids the target distribution from being Gaussian, for instance (for which $||\nabla V(x)||$ is of the order of $||x||$ which is unbounded). It would be interesting to investigate whether this hypothesis could be relaxed, for example, by only requiring $||\nabla V(x)||$ to be bounded on any arbitrarily large compact rather than uniformly on the whole space $\mathcal{X}$, which would allow Gaussian distributions.

Another remark is the recommended h (in the RBF kernel) in the original article [3] that, after testing, does not work well.

In this report, after exploring the inequalities presented in [1], we have studied the behaviour of SVGD when the parameters change. We have observed that SVGD performs excellently in almost all cases (by requiring more or fewer iterations), but the final shape largely depends on the usage. Specifically, when using RBF kernel, it depends on h.

References

[1] Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for stein variational gradient descent, 2021.

[2] Qiang Liu. Liu, q. (2017). stein variational gradient descent as gradient flow. in advances in neural information processing systems, 2017.

[3] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm, 2019.